An Association Rules Based Method for Imputing Missing Ordinal Data and Likert Scales

Proposal for Master's Dissertation

Julian Hatwell

May 5, 2017

Student Number: S15142087
Tutor: Prof. Mohamed Medhat Gaber

Contents

1 Abstract

2 Introduction

This project investigates novel techniques for imputing missing survey data, specifically of ordinal responses and multiple-item (Likert) scales where imputation techniques are not as well developed as they are for continuous data.

2.1 Problem Statement

Surveys using ordinal scales are ubiquitous in the medical and social sciences, market research, election polling and numerous other applications. The practicalities of collecting survey data means that researchers and analysts who use such instruments are regularly faced with the problem of missing data (Bono

et al. 2007; Kamakura and Wedel 2000). Examples are common in the literature (Madow and Olkin 1983; Roth, Switzer III, and Switzer 1999; Raaijmakers 1999) showing that analysis of such survey data is adversely affected by missingness which can result in biased parameter estimates (central tendency, dispersion and correlation coefficients) and loss of statistical power. There are many strategies for recovering from issues caused by missing data which are well developed for continuous variables, but less so for categorical and ordinal variables according to Finch 2010.

There is evidence in the literature of significant differences in the performance of various imputation strategies over different datasets (Wu, Jia, and Enders 2015) *CITE MORE*. This indicates a need to take into account the model and magnitude of missingness and other charactieristics particular to the target data. For this reason there is a real benefit in having a range of imputation methods from which to choose the most suitable for a given situation.

Huisman 1999 states that a strong relationship exists between the items of multiple-item scales which measure one latent trait (such as Likert scales). Techniques that can recognize this information and preserve it in the imputed data set would have an advantage. Association rules are known to exploit a combination of similarity and probabilistic *CITE* features of the underlying distribution which could form the basis of an imputation method. Yet a search of the literature yields no information on the use of association rules in this context.

2.2 Contribution of This Work

This research project will investigate novel, association rules based imputation techniques for missing ordinal and Likert scale data, benchmarked against state of the art. Association rules offer additional advantages of generating compact and highly interpretable models of the target dataset *CITE* which may be used to enhance the results of the analysis.

3 Literature Review

This section provides an in brief review which will be developed further in the dissertation.

3.1 Ordinal Data and Multiple-Item Scales in Surveys

3.2 Models of Missingness of Survey Data

3.3 Imputation Techniques

These include list-wise deletion, regression, stochastic and similarity based methods.

3.4 Benchmarking Imputation Techniques

3.5 Association Rules

3.6 Prediction with Association Rules

References

Bono, Christine et al. (2007). "Missing data on the Center for Epidemiologic Studies Depression Scale: a comparison of 4 imputation techniques". In: *Research in Social and Administrative Pharmacy* 3.1, pp. 1–27.

Finch, W Holmes (2010). "Imputation methods for missing categorical questionnaire data: A comparison of approaches". In: *Journal of Data Science* 8.3, pp. 361–378.

Huisman, Mark (1999). "Missing data in behavioral science research: Investigation of a collection of data sets". In: *M AND T SERIES* 32, pp. 23–46.

Kamakura, Wagner A and Michel Wedel (2000). "Factor analysis and missing data". In: *Journal of Marketing Research* 37.4, pp. 490–498.

Madow, William G and Ingram Olkin (1983). "Incomplete data in sample surveys. Vol. 3: proceedings of the symposium". In:

Raaijmakers, Quinten AW (1999). "Effectiveness of different missing data treatments in surveys with Likert-type data: Introducing the relative mean substitution approach". In: *Educational and Psychological Measurement* 59.5, pp. 725–748.

Roth, Philip L, Fred S Switzer III, and Deborah M Switzer (1999). "Missing data in multiple item scales: A Monte Carlo analysis of missing data techniques". In: *Organizational Research Methods* 2.3, pp. 211–232.

Wu, Wei, Fan Jia, and Craig Enders (2015). "A comparison of imputation strategies for ordinal missing data on Likert scale variables". In: *Multivariate behavioral research* 50.5, pp. 484–503.